Lp-METHODS FOR ROBUST REGRESSION

by
Håkan Ekblom

Lund 1974

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41552412_0033

Lp-METHODS FOR ROBUST REGRESSION

by
Håkan Ekblom

To B.

UNIV.
LEIDEN
BIBL.

This dissertation is based on the following papers:

1. L_p-Criteria for the Estimation of Location Parameters, SIAM Journal on Applied Mathematics, Vol 17, No 6 (1969), 1130-1141. Together with S. Henriksson.

2. A Monte Carlo Investigation of Mode Estimators in Small Samples, Journal of the Royal Statistical Society Series C, Vol 21, No 2 (1972), 177-184.

3. A Note on Nonlinear Median Estimators, Journal of the American Statistical Association, Vol 68, No 324 (1973), 431-432.

4. Calculation of Linear Best L_p-Approximations, BIT 13, No 3 (1973), 292-300.

5. L_p-Methods for Robust Regression.
   To appear in BIT.

1. Introduction

Certainly the most frequently used statistic for a set of observations $f_1, f_2, \dots, f_m$ is the mean $\bar{f} = \sum_{v=1}^{m} f_v/m$. However, this way of concentrating data has certain serious drawbacks which have bothered many people for a long period of time. Thus, if one single observation heavily deviates from the others (which may depend on an observation error or writing or punching error), its influence on the mean value will be unreasonably strong. Huber [6] cites a description of how this problem was delt with more than 150 years ago when the mean yield in certain provinces of France was determined. In that case, when the result from a 20 year period was to be presented, the highest and the lowest values were neglected and averaging was made over the remaining 18 years. This is a technique nowadays called trimming, obviously based upon an ordering of the sample: $f_{(1)} \leq f_{(2)} \leq \dots \leq f_{(m)}$. Of course we can go on and skip a higher fraction of the observations than in the "French example", and ultimately we would reach another well-known statistic, the median:

(1) $$\text{median} = \begin{cases} f_{(k)} & m = 2k-1 \\ (f_{(k)} + f_{(k+1)})/2 & m = 2k. \end{cases}$$

The median illuminates the very nature of the problem. A very strong trimming has made this statistic most insensitive to "outliers" or "wild points"; it is a so called robust estimate. On the other hand, a lot of information in the sample seems to have been thrown away. One natural compromise is to try something between the mean and the median (or their counterparts in more general cases), and this has actually been the origin of many efforts to solve the aforementioned problem.

Example 1. Simple estimation of location.
We have measured some quantity 7 times and recorded the following values: 135,106,103,96,97,83,99. Assume that the "true" value is 100 and that the values 135 and 83 are outliers, i.e. influenced by some "gross error" in addition to the "normal variation". When analyzing the recordings, we do not have that explicit knowledge. We only know

that one or two outliers may be present, and for this reason we use a trimmed mean where the two extreme values are skipped. This gives the value 100.2, to be compared with 99 for the median and 102.7 for the mean.

The mean value, the trimmed means and the median are all unbiased linear estimates for symmetric distributions, i.e. of the form

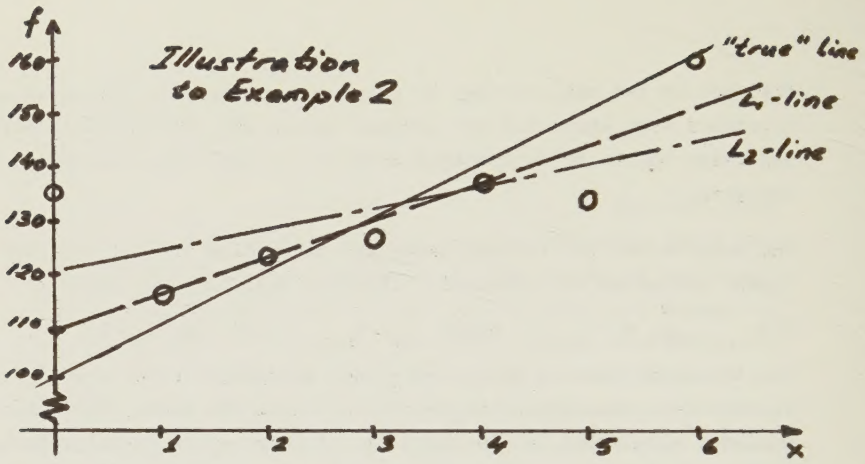
(2) $\sum_{v=1}^{m} \beta_v f_{(v)}$, where $\beta_v = \beta_{m-v+1}$ and $\sum_{v=1}^{m} \beta_v = 1$.

For the case where we have a very good knowledge of the underlying distribution from which the sample is taken, the BLUEs (best linear unbiased estimates) can be computed, at least approximately. These BLUEs are most often highly efficient. The main drawback of linear estimates, however, is that they do not carry over directly to more general situations where no natural ordering of the sample is possible.

Example 2. Straight line regression. (Figure on next page)
Let us assume that the recordings of Example 1 were from the 7 time points $x = 0,1,2,3,4,5$ and 6. Assume further that the result, starting at 100, is "normally" increased by 10 between two recordings. Thus the "true" values are now along a straight line 100+10x. We assume the observations to be influenced by the same errors as before, and thus we now have the following table to analyze:

x	0	1	2	3	4	5	6
f	135	116	123	126	137	133	159

In this new situation, where a straight line is to be fitted to the data points, we can make no reasonable "ordering" (and hence no trimming) of the sample directly. The situation can be classified as slightly paradoxical: only when the "true" straight line is known (or we have a very good approximation to it), we can identify the outlying observations we would like to skip. However, we can compute estimates corresponding to the mean and the median, namely the L_2- and L_1-solutions. For the data given, these lines will be 120+4x and 109+7x. This is a typical result, since L_1 is in general more insensitive to outliers.



2. The robust regression problem

Let us state the problem sketched in Example 2 in a more stringent way. For m values x_ν $(\nu = 1,2,\ldots,m)$ within the interval (x_{min}, x_{max}) we have the observations f_ν $(\nu = 1,2,\ldots,m)$. We suppose that it is reasonable to apply the linear approximation model

$$(3) \qquad f_\nu = \lambda(\alpha, x_\nu) + \epsilon_\nu, \quad \nu = 1,2,\ldots,m$$

where $$\lambda(\alpha,x) = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

Here $\phi_i(x)$ $(i = 1,2,\ldots,n)$ is a base of functions we have chosen and ϵ_ν are random errors with identical distributions. Our task is to compute a function $\lambda(a,x)$ which is a good approximation to $\lambda(\alpha,x)$ in spite of the disturbing errors ϵ_ν present. It should be mentioned that the situation is equivalent to solving the overdetermined system of equations $\lambda(a,x_\nu) = f_\nu$ $(\nu = 1,2,\ldots,m)$. We want to make a comparison between some competing methods to solve this problem. At least for small samples (i.e. small values of m), it is usually impossible to derive any nice analytical results, so we are forced into numerical experimentation through the Monte Carlo technique. For this purpose, we have to define the competition rules.

First let us assume that the error distribution is specified and that we want to test the behavior of a certain method for a model of

type (3). Using a computer we can generate a set of samples and apply the method to each sample. From s samples we will then get s functions $\lambda(a^j,x) = \sum_{i=1}^{n} a_i^j \phi_i(x)$, $(j = 1,2,\ldots,s)$, which are approximations to $\lambda(\alpha,x)$. In order to rank the methods, we have to produce one single "goodness" figure for each method.
We will use the Mean Integrated Square Error

$$(4) \qquad MISE = \frac{1}{s} \sum_{j=1}^{s} \int_{x_{min}}^{x_{max}} w(x)\,(\lambda(\alpha,x) - \lambda(a^j,x))^2 dx$$

Now, if we have chosen the $\phi_i(x)$ to form an orthonormal set of functions on (x_{min},x_{max}) with respect to $w(x)$, (4) will reduce to the sum of the estimated Mean Square Errors:

$$(5) \qquad MISE = \sum_{i=1}^{n} MSE_i\,, \qquad MSE_i = \sum_{j=1}^{s} (a_i^j - \alpha_i)^2/s.$$

In the simplest case, n=1 and $\phi_1(x)=1$, which is usually called simple estimation of location, a slightly different measure has also been used. If we have specified the error distribution to be symmetrical, we can use the variance of a_1 instead of MSE_1. This is done in Papers 1, 2 and 3. The standard errors of the MSEs or the variances will then give an idea of the precision of the results. In Papers 1 and 2 attempts instead are made to construct confidence intervals for these dispersion measures.

Once we have adopted a "goodness" measure, e.g. the MISE, we can apply different methods to the same set of samples from a distribution and get a ranking of the methods. Generally we are interested in the results from not only one single distribution but rather a whole family of error distributions. In the Papers 1,2,3 and 5 the Monte Carlo experiment utilizes such families with 4-8 members.

The Normal, Laplace and Cauchy distributions, with frequency functions essentially of the form $\exp(-x^2)$, $\exp(-\text{abs}(x))$ and $1/(1+x^2)$, respectively, are most often present in investigations of this kind. In the order they are mentioned, they are characterized by an increasing "thickness" in their tails.

The classical way of describing the outlier problem is through a model with contaminated normal distributions. Thus we can take 90% of the observations from an N(0,1) (i.e. normal distribution with zero mean and standard deviation 1) and 10% from an N(0,4) in order to represent the outliers. Of course the contamination need not be symmetrical. E.g. in Paper 5, one type of samples is built up from 19 N(0,1)-observations and 2 values from N(4,1).

3. L_p-methods

In the approximation situation described at the beginning of the previous section, let us introduce the residuals $r_v(a) = f_v - \lambda(a, x_v)$, $v = 1,2,\ldots,m$. For the L_p-methods, the following quantity is minimized with respect to a:

$$(6) \qquad \left(\sum_{v=1}^{m} |r_v(a)|^p\right)^{1/p}$$

For $p \geq 1$, this is a vector norm. As is well known, p=2 (least squares), $p=\infty$ (Chebyshev approximation) and to some extent p=1 ("approximation in the mean") constitute the established L_p-methods. However, during the last decade an increasing interest in other p-values has become apparent.

In his thesis, Gentleman [3] considered L_p-methods with $1<p<2$ for multi-dimensional point estimation. Rice and White [8] have published a comparison between some L_p-methods for $p \geq 1$. Essentially for estimation of location, the statistical efficiencies were computed for a family of error distributions and for varying sample sizes. It should be pointed out that the small sample results are the more interesting, since the asymptotic efficiencies (i.e. for $m=\infty$) can be derived from Huber's results [5]. Paper 1 is built upon this work of Rice and White, but there is an extension of the L_p-norm (for the simple estimation of location case) to an L_p-criterion for all p. E.g., for $p=-\infty$ we are led to minimization of $(f_{(v)}-f_{(v-1)})$ over v, while for p=0 we seek the point f_v which makes $\prod_{j \neq v} |f_j - f_v|$ a minimum.

It is obvious that the more the p-value is decreased from 1, the more the cluster points of the sample will be favorized in our search for the L_p-minimum. For $p=-\infty$ we have come to an extreme situation where the two closest points totally determine the estimate. This estimator is quite similar to proposals for finding the mode of a distribution. A natural question then arises: how do the L_p-methods for $p \leq 1$ behave compared to other mode estimators? Paper 2 is an attempt to give an answer to that question, but for economical reasons the Monte Carlo investigation only considers p=1, 0 and $-\infty$ for the L_p-methods. Broadly speaking, the result is that L_0 is almost as efficient as the best of the non-L_p-minimizers. In addition to this result, it should be stressed that L_p-methods have two advantages: 1) once the p-value is fixed we do not have to bother about choosing (sometimes mysterious) parameters, 2) the definition is readily extended to more general estimation cases.

One reason why p=1, 2 and ∞ have become the popular values for L_p-minimization is that the computation is simple, compared with the amount of work required for other p-values. This is equally true for the simple estimation of location and raises the question if the non-standard L_p-methods could be simplified while still keeping their estimating properties. Paper 3 presents two median-like simple estimates based upon this principle. One of them is better than the traditional median for a large family of distributions, while the other is applicable to the Cauchy distribution.

The general result of Papers 1-3 is that L_p-methods can compete with other one-dimensional point estimators but will win only in quite special situations. It should be mentioned that a very thorough study of other possibilities is given in [1].

Papers 1-3 only consider the simplest case, namely simple estimation of location. There are two reasons for this: 1) the lack of efficient algorithms for the case n>1 would have made the computation very costly, and 2) there is widespread belief that the case n=1 tells us very much about the statistical efficiency of the methods in general. Concerning the

first point, the situation has recently improved through the work by Barrodale-Roberts [2] (for p=1) and Henriksson [4] (for $0 < p < 1$) and also through the algorithm proposed in Paper 4 for $1 < p < 2$. The latter algorithm is based upon the damped Newton method which, however, is not directly applicable due to the possibility of infinite elements in the Hessian matrix. To overcome these troubles, a sequence of perturbed problems is constructed to which the Newton iteration can be applied. With this imbedding technique, we can achieve that the iteration in each perturbed problem rapidly becomes quadratically convergent.

In Paper 5 a very simple example is given which shows that for the special case n=2, m=3, the L_2-method is superior to the L_1-method for, e.g., Laplace distributed errors. This result is in contrast to the situation for n=1. Since that example contradicts the belief of point 2 above, it is a motivation for a comparison between methods for the straight line and parabolic regression case (n=2 and 3, m=21) as presented in Paper 5. The L_p-methods for p=2, 1.5, 1.25, 1 and 0.75 are all included as well as an estimate proposed by Huber [7], which uses the L_2-criterion for small residuals and L_1 for larger ones. Although the precision in the Monte Carlo results is not overwhelming, it seems quite clear that the L_1-method is less efficient than the results for n=1 indicate. The Huber estimate is certainly the best choice for contaminated normal errors, but for a larger family of distributions the $L_{1.25}$-method is equally attractive. For error distributions with very long "tails" (and perhaps also for strongly asymmetric ones), L_1 or $L_{0.75}$ is to be preferred. Finally, it must be stressed that the least squares criterion is a dangerous choice if we do not have a normal error distribution.

4. Some final remarks

There are many details in the L_p-algorithm of Paper 4 which need further consideration, until an efficient implementation is available. Further, a fair comparison with other possible methods for computing L_p-solutions, especially for $1<p<2$ but also for $2<p<\infty$, is highly desirable. Such an evaluation is definitely a non-trivial task. Actually,

it has been pointed out (e.g. by J.N. Lyness at the IFIP Congress 1971) that it is most often an order of magnitude easier to write two sophisticated routines than to determine which of the two is better.

Similarly, the Monte Carlo results have raised new questions, e.g. concerning the influence of the sample size (m) and the distribution of the arguments (x_v) on the results and also concerning the choice of parameter value and scaling technique in the Huber estimate. This situation is drastically summarized by John von Neumann's paraphrase, which has been interpreted by Tukey [9]: " 'The only good Monte Carlo is a dead Monte Carlo'. This aphorism was coined to express the view that out of a well-conducted Monte Carlo should come enough insight to allow us to use newly-developed or newly-chosen approximations to solve other cases of that particular complexity, thus needing to use Monte Carlo again only when we want to go still further or still deeper. I approve this goal; I only wish I could reach it more often."

Acknowledgements

First I would like to thank Professor Carl-Erik Fröberg, my teacher, who has furnished encouragement and helpful criticism, and has given me particularly valuable advice on the presentation of the results.

The work presented in Papers 1-3 was carried out together with Dr Sten Henriksson. Also to the last two papers, he has contributed numerous ideas and advice. This cooperation has been immensely valuable to me.

Many people working in computer science and related fields have offered suggestions and constructive criticism. Dr Per-Åke Wedin has provided continuous support during the last year. I also want to thank, among others, Professors E.W. Cheney, J.R. Rice, P.J. Huber, I. Barrodale, Dr Jan Lanke and Dr Axel Ruhe for helpful comments.

Most of the research presented was carried out during my time at the Department of Computer Sciences at Lund University. These years offered many interesting discussions with my colleagues on a lot of subjects. During the second half of the dissertation work (Papers 4 and 5), I have been employed at the Computing Center of Lund University. The confrontation with the more practical aspects of computing, together with the very nice working environment as a whole, made this a very stimulating period. Further, I am most grateful to Ingemar Dahlstrand and Lennart Bensryd for giving me all the free time I asked for. Also my present employer, University of Luleå, has provided very good working conditions during this fall.

Finally, I wish to thank Miss Ingegerd Nyberg for efficiently typing and patiently retyping this manuscript and my children, Annika, Petra and Pontus, for producing a stimulating background noise.

References

1. D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers and J.W. Tukey, Robust Estimates of Location: Survey and Advances. Princeton University Press (1972).

2. I. Barrodale and F.D.K. Roberts, Solution of an Overdetermined System of Equations in the L_1 Norm, Mathematical Dept Report No 69, University of Victoria (1972).

3. W.M. Gentleman, Robust Estimation of Multivariate Location by Minimizing pth Power Deviations, Unpublished Ph.D. Thesis, Princeton University, Princeton, New Jersey (1965).

4. S. Henriksson, On a Generalization of L_p-Approximation and Estimation, Thesis, Dept of Computer Sciences, Lund University, Sweden (1972).

5. P.J. Huber, Robust Estimation of a Location Parameter, Annals of Mathematical Statistics 35 (1964), 73-101.

6. P.J. Huber, Robust Statistics: A Review, Annals of Mathematical Statistics 43, No 4 (1972), 1041-1067.

7. P.J. Huber, Robust Regression, Paper presented as part of the Wald Lecture at the IMS Annual Meeting at Hannover, New Hampshire (1972).

8. J.R. Rice and J.S. White, Norms for Smoothing and Estimation, SIAM Review 6 (1964), 243-256.

9. J.W. Tukey, Data Analysis, Computation and Mathematics, Quarterly of Applied Mathematics 30 (1972), 51-65.

AV-centralen, Lund 1973